Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 51-32 (2018) 55–60

Optimization of Sentiment Analysis Methods for classifying text comments of bank customers

M. Lutfullaeva*
M. Medvedeva, Candidate of Physic-Mathematical Sciences, Associate Professor **
E. Komotskiy***
K. Spasov, PhD****

* Ural Federal University. The First President Of Russia B. N. Yeltsin, Yekaterinburg, Russia (e-mail: malikalut!704@smail.com)
** *Ural Federal University. The First President Of Russia B. N. Yeltsin, Yekaterinburg, Russia (e-mail:marmed55@yandex.ru)*
*** *Ural Federal University. The First President Of Russia B. N. Yeltsin, Yekaterinburg, Russia (e-mail:ekomotskiy@yandex.ru)*
**** *Sofia University "St.Kliment Ohridski",Sofia, Bulgaria (e-mail:kspasov@fmi.uni-sofia.bg)*

Abstract: A method of sentiment analysis of the text and its approbation in solving the problem of analysis of text comments left by the Bank's customers are performed. The proposed method consists in a combination of three approaches: rules-based, dictionaries and machine learning with a teacher. New method of text vectorization- tonal vectorization instead of classical ones, such as "bag-of-words " and TF-IDF, is proposed. The text was classified by logistic regression with regularization. A series of experiments were carried out and the optimal value of the regularization parameter was found in terms of classification accuracy.

Keywords: sentiment analysis, sentiment of the text, tonal dictionary, tonal vectorizer, the bag-of-words, machine learning , optimization

1. INTRODUCTION

In recent years, the problem of text classification by emotional colour is popular in scientific community, which is also known as sentiment analysis. Sentiment analysis is part of computational linguistics that studies opinions and emotions in text documents and is a set of methods designed to detect emotions reaction or attitude (sentiment) expressed in the text automatically.

This analysis allows to find out the emotional colour of the text, which could be positive, negative or neutral [2]. Sentiment of the text is determined by the lexical key of its constituent units and the rules of their combination [1]. Sentiment of the text is determined by three factors: the subject of tonality, the tonal evaluation, the object of tonality. The subject of tonality is the author of the text; the object of tonality is what or who this text is about [2]. Tonal estimation can be presented in one of the following types: binary (positive / negative), ternary (positive / neutral / negative), ranked [3]. Existing methods and algorithms are unable to show sufficiently high results in determining the sentiment of the text, because of complexity of the problem of automatic determination of emotional colour of the text. Therefore there is a need to improve the technique.

The purpose of this article is to improve methods of sentiment analysis of the text and to test it in task of making opinion mining of text comments left by customers of the Banks. The solution of current problem will allow to understand when the customers of the Banks are satisfied and dissatisfied, what are the problems of service and to understand customers' the attitude to the Bank. For the banking sector the extraction of such information is important since this sphere presence of a large number of competitors. In this case banks requires the creation of the most attractive service conditions.

Existing approaches to sentiment analysis are divided into several categories (table 1).

Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2018.11.353

Table 1 Existing approaches to sentiment analysis

№	Approaches to sentiment analysis	Description of the approach group
1	Based on the rules	The essence of the approaches of this group presents by set of rules, according to them the system makes a conclusion about the tonality of the text (for example, the presence of the word " not "before the word "good", bearing a positive character, changes the character of the phrase to negative - " not good»);
2	Based on the dictionaries	The main idea is to compile so-called tonal dictionaries, that is sets of words and emotional states that have a numerical positive or negative tonal evaluation
3	Supervised machine learning	The essence of the algorithms is to train the classifier using labeled data set (collection of texts), and then use the resulting model for the analysis of new text documents.
4	Unsupervised machine learning	Those algorithms are based on the machine learning classifier, which isn't trained on previously labeled data set, which implies that the algorithm will find hidden patterns in the data then, according that, divide documents into groups.

Researchers, who solves problems of computational linguistics, including the problems of text classification, often chose one approach - machine learning. This approach gives good results in the thematic classification of the text (classification by topic, genre, etc.), as it is based on the frequency of word analysis and identification of unique words that characterize a certain group of texts, for example, identifying political articles by keywords. However, this approach is not suitable for sentiment analysis, as the words that shows emotional states are not unique for text document, and therefore poorly taken into account by the algorithm of machinery training when using a standard texts vectorizers, such as "bag-of-words" and TF-IDF (term frequency-inverse document frequency).

2. METHODS

Due to the disadvantages of the using of standard approaches to sentiment analysis, it was proposed to use a technique that is based in synthesis of three approaches: rules-based, dictionaries and supervised machine learning. In addition, it is proposed to use a different approach to text vectorization - tonal vectorization, different from standard methods such as "bag-of- words" and TF-IDF. The main idea is to find in the text comment the words, which express positive or negative assessment, and based on this analysis, evaluate sentiment of that comment - positive or negative. The proposed method involves several stages in the implementation of the sentiment analysis (figure 1).

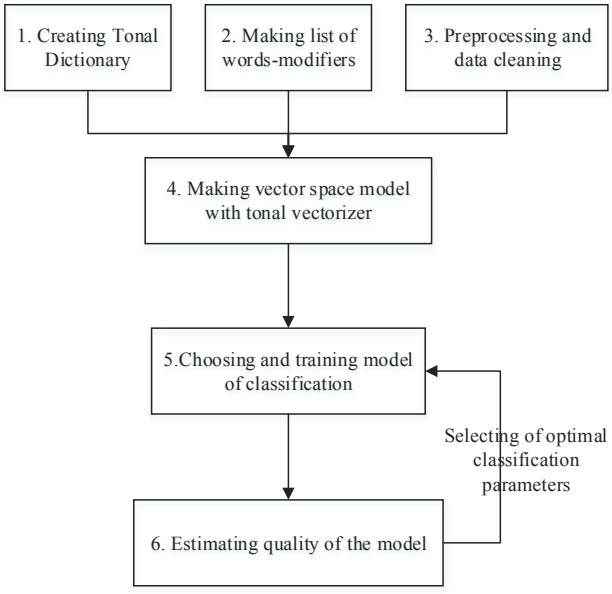

Fig. 1 Stages of sentiment analysis

Creating a tonal dictionary involves the selection of words that express positive or negative attitude (emotional states), and grouping them by the nature and extent of the sentiment. Each group of tonal words is assigned a numerical value that conveys the sentiment. For these purposes, it is easy to use a special scale, according to which negative tonal estimates are matched with negative numbers, and positive - positive (for example, from -3 to 3).

The specificity of the analysis of customer comments is that the most interesting are clearly positive or clearly negative reviews, and therefore texts with neutral text information do not need to be processed. That's why in the scale for the assessment of the tone of the words is not used a value of zero. The scale scope depends on the required degree of accuracy of the sentiment analysis: the more detailed analysis is needed, the wider the scale should be. Table 2 shows an example of a tonal dictionary with a scale from -3 to 3

Table 2. Groups of words by emotional characteristics in the tonal dictionary

Grade	Description of emotional characteristics	Examples of words in the dictionary
3	Extremely positive emotional characteristic. Expresses absolute customer satisfaction with the service/services	Great, perfect, fantastic, super, perfect
2	Positive evaluation of the service / services by the client	good, affordable, intelligible, pleasant, responsible, responsive
1	Weakly expressed in positive customer satisfaction, partial satisfaction with the service/services	okay, not bad, well, averagely
-1	Weakly expressed negative evaluation of customers, partial satisfaction with service/services	so-so, mediocre, weak, it was better
-2	Negative evaluation of the service / services by the client	bad, unclear, unpleasant, pity, annoying, unprofitable
-3	Extremely negative emotional characteristic. Expresses absolute dissatisfaction with the customer service/services	horrible, outrageous, disgusting, nightmare, worse than ever

The next step is to determine the emotional states (words-modifiers), the nature and extent of their influence on the existing emotional characteristic of the text. Often, the natural languages have the linguistic constructions which could strengthen or weaken or even change the sentiment of the whole expression. This stage involves the definition of such words constructions and assigning them numerical values that correct the words from the tonal dictionary defined at the previous stage.

There are two main groups of words-modifiers:

- which increase emotional assessment (words like "very", "fairly", "highly", "very", "even", etc.);

- which decrease emotional assessment ("happened"," not really"," not really", etc.);

For example, if the word "good" refers to a positive assessment, the word-modifier "very" increases the positive assessment, the result is "very good". In turn, if you put the word "not" before this phrase, we get a restrained negative assessment of "not very good". That is way it is important to take into account the influence of words-modifiers on the emotional colour of the review.

Next step is text pre-processing and cleaning. This stage is necessary step to make analysis easier. Classically, the following steps are related to pre-processing and cleaning of the text: splitting the text into semantic units (tokenization), stemming and lemmatization, removing stop words, clearing from punctuation marks.

The next step is to convert cleared text into the vector of parameters, which will be fed to the classifier input. This process is called vectorization. Within the framework of this work, the authors propose to carry out the so-called tonal vectorization. In contrast to classical methods, such as a “bag-of- words” or TF-IDF, which reflect the frequency of words entering the document, so that all significant words in the document act as signs [4]. This approach works well with thematic classification, in which unique, specialized on the topic of the word can identify the theme of the entire text and then carry out the classification of topics. However, using of these methods, words that convey emotional states are not defined as sufficiently significant, because they are not "rare" for a particular comment on the entire array of comments.

Because of mentioned problem, it became necessary to develop and use a new approach. Its essence is to carry out the vectorization of the analysed comments, using a tonal dictionary compiled at the first stage. The length of the feature vector corresponds to the number of words in the tonal dictionary, each word of which is have personal sign. If a word from the tonal dictionary is present in the text (comment), then for this comment a numerical assessment of the emotional characteristic of the word is entered into the feature vector. This approach allows to monitor the presence of emotionally significant words in the text and, accordingly, allows to get a numerical assessment of the tone of the comment.

Further, when feature vectors are formed, the binary classifier is trained based on the training sample. The quality of its work is carried out on a test sample. In this work, logistic regression with regularization was used as a classification model. The logistic regression equation in general has the following form:

$$f(z) = \frac{1}{1 + e^{-z}} \quad (1)$$

z - vectors-columns of independent variables values x (feature vector) and values of parameters (weights) w for n variables:

$$z = w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n \quad (2)$$

The use of regularization allows to configure the training of the classifier so that, on the one hand, to avoid retraining, and on the other to achieve the maximum possible accuracy of the classification.

The estimation of classifier accuracy is calculated by the following formula:

$$Accurancy = \frac{N}{P} \cdot 100\% \quad (3)$$

N - number of correctly classified documents;
P - number of documents in the training sample.

3.RESULTS

For carrying sentiment analysis according to the described method, we made sample consisting of 20 000 text comments of the Bank's clients. The ratio of training and test samples is 80/20.

Before the analysis, a tonal dictionary was formed, it contains words that convey positive and negative emotional color of the three levels of expression. Each word from dictionary is rated on a scale from -3 to 3. Table 2 presents a description of each group of tonal words by the degree of expression of positive and negative attitudes, as well as examples of words that were put into different groups.

Pre-processing and cleaning the original comments included the following steps:

- bringing words to lower register;

- deleting punctuation marks;

- removing stop-words;

- stemming;

Further, the cleared text of comments was translated into feature vectors, on the basis of which the logistic regression model was trained. We compared the performance of the classifier at different vectorizers, we used "bag-of-words", TF-IDF vectorizer and the method, which were described in this work. Table 3 presents examples of logistic regression work (we founded the probability that logistic regression reveals the correct emotional characteristic) with classifying some comments using different vectors.

Table 3. Results of classification of logistic regression using different vectors

The text of the comment	Characteristic	Logistic regression result for different vectors		
		«word bag»	TF-IDF	Tonal vectorization
“I was faced with an absolutely boorish, disdainful attitude of the staff in the bank office”.	negative	0,22	0,31	0, 97
“The reason of two-month production of the certificate on payments distinctly can't explain. Very unpleasant with quality of service”	negative	0,14	0,18	0,92
“The operator for any questions writes the details without the snobbery and any sign of tiredness. She does her job well and professionally. I am satisfied with the service”	positive	0,31	0,27	0,96
“Well done, that and say. Work promptly and professionally.”	positive	0,26	0,35	0,97
“Spent today 2 minutes to pay. Previously, had to spend from 15 minutes to 1 hour”	positive	0,15	0,18	0,17
“They promised to contact me on my issue in the coming three days. It's been three weeks. This is the attitude of the Bank's employees.”	negative	0, 17	0, 15	0, 16

As can be seen from the table, when using a tonal vectorizer, the logistic regression produces significantly better results and more accurately reveals the emotional characteristic of the text. This is happening due to the fact that the feature vector given to the input of the classifier contains information about the presence or absence of words from the tonal dictionary. However, this method does not work well when the text does not contain explicit positive and negative expressions, as, for example, in the last two comments from the table. In the last two cases, the negative and positive attitude of clients was manifested without the use of words from the tonal dictionary, the emotional characteristic of such comments was transmitted by the context. In such cases, the most difficult to carry out a sentiment analysis, as the

vectorizator and the classifier can not catch the meaning of the sentence, unlike a person.

Further, a series of experiments on the classifier model was carried out, which allowed to identify such an optimal value of the logistic regression regularization parameter (C), in which the accuracy of the classification results would be maximized. In order to conduct a comparative analysis, training and verification of the accuracy of the classifier on the basis of features formed with the help of different classifiers were conducted. Figure 2 shows a graph of the accuracy of the classifier depending on the values of the regularization parameter C. the Accuracy is calculated by the formula (3).

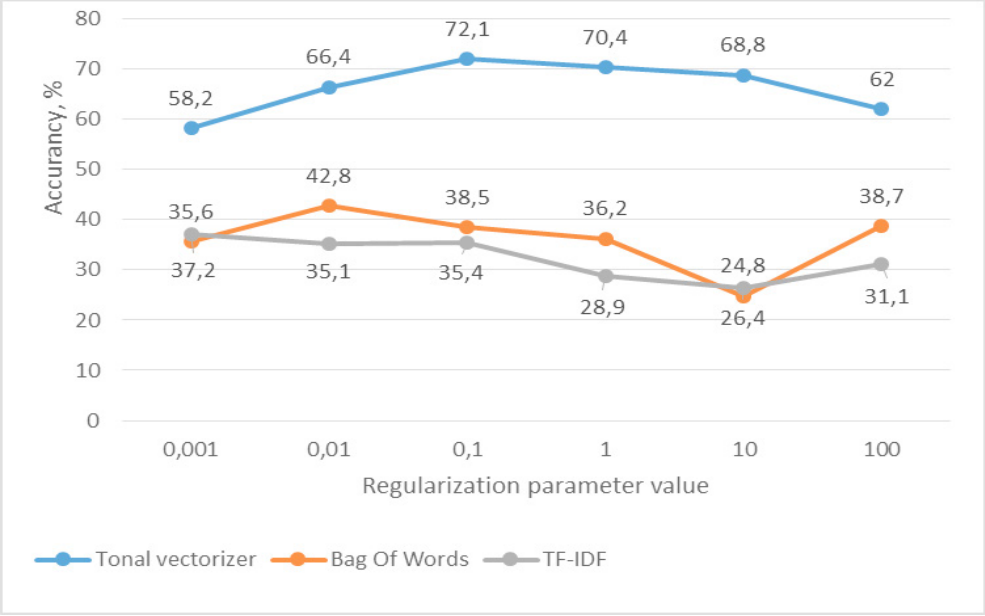

Fig.2. The dependence of classification accuracy on the regularization parameter C for different vectors

As can be seen from picture 2, the classification of text using a tonal vectorizer shows much better results. At the same time, during the run of the model, it was revealed that the classification is carried out with a maximum accuracy of 72.1% with the regularization parameter C = 0.1.

4. SUMMARY

In this paper, we propose a method of sentiment analysis of the text and its approbation in solving the problem of analysis of text comments left by the Bank's customers. During the research we analyzed the feasibility of using classical vectorizer ("bag-of-words", TF-IDF) while using conducting sentiment analysis. It is revealed that these methods are not suitable for solving the problem of revealing the emotional characteristics of the text. In this regard, it is proposed to use a tonal vector and conducted a series of experiments to confirm/refute the effectiveness of this approach. The results obtained show that tonal vectorization, in contrast to classical vectorizers, allows to determine unambiguously the tone of an obviously emotionally-colored comment (92-97%).

The search for the optimal regularization parameter (C) was carried out. The classification accuracy for different parameters with varied from 58.2% to 72.1%, the maximum accuracy achieved C = 0.1. While using the word bag and TF-IDF, the maximum classification accuracy was only 42.8%.

It was also revealed that the proposed approach to sentimentalize bad classified texts without strong emotional tone, because it is focused on the search words from the tone vocabulary. The solution to this problem requires further research.

REFERENCES

1. Nugumanova A., Bessmertnyi I. Applying the latent semantic analysis to the issue of automatic extraction of collocations from the domain texts Please use this document as a “template” to prepare your manuscript. For submission guidelines, follow instructions on paper submission system as well as the Conference website. // Communications in Computer and Information Science. 2013. V. 394. P. 92-101. doi: 10.1007/978-3-642-41360- 5_8

2. Cruz F.L., Troyano J.A., Pontes B., Ortega F.J. Building layered, multilingual sentiment lexicons at synset and lemma levels // Expert Systems with Applications. 2014. V. 41. N 13. P. 5984-5994. doi: 10.1016/j.eswa.2014.04.005

3. Parau P., Stef A., Lemnaru C., Dinsoreanu M., Potolea R. Using community detection for sentiment analysis // Proc. IEEE 9th Int. Conf. on Intelligent Computer Communication and Processing (ICCP 2013). 2013. P. 51-54/ doi:10.1109/ICCP.2013.6646080

4. Chiru C.-G., Hadgu A.T. Sentiment- based text segmentation // Proc. 2nd Int. Conf. on Systems and Computer Science (ICSCS 2013). 2013. P. 234-239. doi: 10.1109/IcConSCS.2013.6632053

5. Esuli A., Sebastiani F. Determining the Semantic Orientation of Terms through Gloss Classification // Conference of Information and Knowledge Management (Bremen). ACM, New York, NY, 2005, pp. 617-624.